(19)대한민국특허청(KR)
(12) 등록특허공보(B1)

(51) 。Int. Cl.[7]
G06F 17/27

(45) 공고일자 2004년02월25일
(11) 등록번호 10-0420096
(24) 등록일자 2004년02월12일

(21) 출원번호 10-2001-0012318
(22) 출원일자 2001년03월09일

(65) 공개번호 10-2002-0072140
(43) 공개일자 2002년09월14일

(73) 특허권자 주식회사 다이퀘스트
서울특별시 서초구 서초동 1604-22 신도빌딩

(72) 발명자 서정연
서울 서초구 반포1동 32-8 삼호가든아파트1차 B동 매슌 203호

이근배
서울 광진구 구의1동 251-103

고영중
경기도 고양시 일산구 주엽동 강선마을 711-902

(74) 대리인 특허법인세신

심사관 : 김승조

(54) 각 범주의 핵심어와 문장간 유사도 측정 기법을 이용한비지도 학습을 기반으로 하는 자동 문서 범주화 방법

요약

본 발명은 수집된 문서를 문장 단위로 나눈 후 각 범주의 핵심어 입력과 문장간 유사도 측정 기법을 사용하여 문장들을 각 범주별로 분류하고 범주별로 모아진 문장들을 학습 데이터로 사용하여 학습하고 문서 범주화 작업을 수행하는 자동 문서 범주화 시스템 및 방법을 제공한다. 본 발명은 (i)수집된 문서를 문장 단위로 분할하고 형태소 분석하여 내용어를 추출하는 단계; (ii)입력된 핵심어를 이용하여 각 범주의 대표 문장을 추출하는 단계; (iii)상기 추출된 대표 문장이 각 범주의 특성을 잘 나타내고 있는지를 검증하여 순위화하는 단계; (iv)상기 추출된 대표 문장과 대표 문장으로 추출되지 못한 미 분류 문장과의 문장간 유사도 측정 기법을 이용하여 학습에 사용될 학습 문장 집합을 생성하는 단계; (v)상기 생성된 학습 문장 집합을 사용하여 자질을 추출하고 학습하여 문서 범주화를 수행하는 단계를 포함하는 비지도 방식의 자동 문서 범주화 방법이다.

대표도

도 1

색인어

자동 문서 범주화, 비지도 학습, 핵심어, 문장간 유사도 측정, 문서 분류기

명세서

도면의 간단한 설명

도 1은 본 발명에 따른 자동 문서 범주화 방법을 나타낸 전체 흐름도.
도 2는 도 1의 전처리 단계에서 내용어 추출 과정의 일 예를 나타낸 흐름도.
도 3은 도 1의 학습 문장 집합 생성 단계의 일 예를 나타낸 흐름도.
도 4는 도 3에서 단어-문장간 유사도 측정의 반복 계산 예시도.

발명의 상세한 설명

발명의 목적

발명이 속하는 기술 및 그 분야의 종래기술

본 발명은 온 라인 상 문서의 자동 문서 범주화에 관한 것이며, 특히 수작업으로 수행되는 대량의 학습 문서 생성 작업 없이 적은 비용으로 문서 범주화를 수행할 수 있는 방법에 관한 것이다.
최근에는 인터넷이 폭 넓게 보급되어 온 라인(on-line)상에서 얻을 수 있는 텍스트(text) 정보의 양이 급증함에 따라 텍스트 문서를 수집하는 것은 쉬워졌으나 수집된 텍스트 정보에 대한 효율적인 정보 관리가 요구되고 있다.
종래의 자동 문서 범주화 방법은 보통 수작업에 의해 범주가 할당된 대량의 학습 문서를 사용해서 학습하고 범주화 작업을 수행한다. 그러나, 학습에 사용될 대량의 양질의 학습 문서를 생성하는데는 많이 비용과 어려움이 있다. 특히, 자동 문서 범주화의 영역이 신문 기사, 전자 도서관뿐만 아니라 전자 우편, 뉴스 그룹 등 적용 영역이 넓어지고 다양해지고 있으므로 각 영역에 따라 대량의 학습 문서를 생성한다는 것은 많은 작업 인원과 많은 시간을 필요로 하는 어려움이 있다.

발명이 이루고자 하는 기술적 과제

상기와 같은 문제점을 해결하기 위한 본 발명의 목적은 학습 문서를 생성하기 위한 작업 없이 각 범주의 핵심어의 입력만으로 인터넷에서 수집된 문서를 사용하여 자동으로 학습 데이터를 생성하고 학습하여 문서 범주화를 수행하는 방법을 제공하는 데 있다.
본 발명은 기본적으로 텍스트 문서를 문장 단위로 나누는 기술과 형태소 분석 및 태깅 기술을 이용하고 있으며, 입력된 핵심어로부터 분류하고자 하는 각 범주의 특징을 잘 내포하고 있는 문장을 자동으로 추출하고 순위화하는 통계적 정보 검색 기법을 사용한다. 또한, 문장간 유사도 측정기법을 이용하여 학습 문장 데이터를 자동으로 구축하기 위한 통계적 언어 분석 기법을 사용하고 있으며, 구축된 학습 문장 데이터를 사용하여 자질을 추출하고 분류하는 과정에 의해 문서 범주화를 이룩한다.

발명의 구성 및 작용

상기한 바와 같은 목적을 달성하기 위하여, 본 발명은 (i)수집된 문서를 문 장 단위로 분할하고 형태소 분석하여 내용어를 추출하는 단계; (ii)입력된 핵심어를 이용하여 각 범주의 대표 문장을 추출하는 단계; (iii)상기 추출된 대표 문장이 각 범주의 특성을 잘 나타내고 있는지를 검증하여 순위화하는 단계; (iv)상기 추출된 대표 문장과 대표 문장으로 추출되지 못한 미 분류 문장과의 문장간 유사도 측정 기법을 이용하여 학습에 사용될 학습 문장 집합을 생성하는 단계; (v)상기 생성된 학습 문장 집합을 사용하여 자질을 추출하고 학습하여 문서 범주화를 수행하는 단계를 포함하는 것을 특징으로 한다.
이하, 첨부된 도면을 참조하여 본 발명을 상세히 설명하기로 한다.
도 1은 본 발명에 따른 학습 문서의 생성 작업없이 각 범주의 핵심어의 입력만으로 수집된 문서를 자동으로 분류해내기 위한 비지도(非指導) 학습 기반의 자동 문서 범주화 방법의 전체 흐름도이다.
도시된 바와 같이, 본 발명의 방법은 전체로 보아 수집된 문서의 형태를 정규화하고 문장 단위로 분할하며 언어적 분석을 통해 각 문장의 내용어를 추출하는 전처리 단계(10); 상기 가공된 문장 집합에서 대표 문장 추출과 문장간 유사도 측정 과정을 거쳐 학습 문장 집합을 자동으로 생성하는 학습 문장 집합 생성 단계(20); 상기 생성된 학습 문장 집합을 사용하여 자질을 추출하고 학습하여 입력 문서를 분류하는 자질 추출 및 범주화 단계(30)로 이루어진다.
상기 전처리 단계(10)는 수집된 문서를 본 시스템에서 사용하기 위해서 기계적 처리가 가능하도록 변환하는 문서 정규화 과정(110)과; 문장 단위 분할 과정(120)과; 형태소 분석 및 태깅 과정(130)과; 문장의 내용이나 특징을 잘 반영하는 내용어를 추출하는 내용어 추출 과정(140)을 포함한다.
문서 정규화 과정(110)은 HTML 문서 등에서 나타나는 태그(tag)와 특수 문자를 제거하고 한자어는 해당하는 한글로 변환시키는 작업을 수행한다.
문장 단위 분할 과정(120)은 한국어의 특징에 맞추어 종료형 어미(~다, ~까, ~요, ~죠 등) 다음에 마침표(.), 물음표(?), 또는 느낌표(!)가 나오는 경우를 문장의 끝으로 보고 문서의 내용을 문장 단위로 분리한다.

형태소 분석 및 태깅 과정(130)은 문장을 언어적, 통계적 분석을 통하여 각 형태소 별로 나누어 품사를 결정한다.
내용어 추출 과정(140)은 문장의 특징을 잘 나타내는 품사인 명사와 동사를 대상으로 문장의 내용어를 추출하는데 명사나 동사 중에도 문장의 내용을 구별하는데 별다른 정보를 주지 못하는 불용어를 처리하기 위해 불용어 사전을 사용하여 불용어 사전에 등록된 단어는 내용어 추출에서 제외된다. 도 2를 참고로 내용어 추출과정에 대한 예를 후술한다.
상기 학습 문장 집합 생성 단계(20)는 입력된 핵심어를 이용하여 각 범주의 대표 문장을 추출하는 대표 문장 추출 과정(210)과; 추출된 대표 문장이 각 범주의 특성을 잘 나타내고 있는지를 검증하여 순위화하는 대표 문장 검증 과정(220)과; 문장간 유사도 측정 기법을 이용하여 최종적인 학습 문장 집합을 생성하는 문장간 유사도 측정 과정(230)을 포함한다. 학습 문장 집합 생성 단계의 일 예를 도 3을 참고로 후술한다.
대표 문장 추출 과정(210)은 입력된 범주별 핵심어를 직접 문장의 핵심어로 가지고 있는 문장들을 추출하여 이들을 각 범주의 특성을 가장 잘 나타내는 문장으로 간주한다.
대표 문장 검증 과정(220)은 핵심어를 포함하고 있는 문장 중에 그 범주에 해당하지 않는 문장이거나 혹은 그 범주의 특성을 잘 나타내지 못하는 문장들을 제거하기 위해서 추출된 문장들을 각 범주의 특성을 잘 나타내는 순위로 순위화하기 위해 문장 가중치를 계산하고 순위화한다. 추출된 대표 문장의 각 내용어에 가중치를 부여하기 위하여 정보 검색 분야에 널리 사용되고 있는 용어 빈도(TF: Term Frequency)와 역범주 빈도(ICF: Inverse Category Frequency)를 사용했으며 문장의 가중치는 계산된 내용어 가중치의 평균값을 사용한다.
문장간 유사도 측정 과정(230)에서 추출된 대표 문장 집합은 문서 범주화의 학습 데이터로 사용하기 위해서는 그 양이 아직 적기 때문에 대표 문장으로 추출되지 못한 문장들을 각 범주의 대표 문장들과의 유사도 측정을 통해 측정된 유사도가 가장 높은 범주에 할당함으로써 학습 문장 집합을 생성한다. 본 발명에서는 단어 유사도 행렬과 문장 유사도 행렬을 사용하여 반복 계산을 통해 문장간 유사도를 계산하는데 그 예는 도 4에서 도식화하였다.
자질 추출 및 범주화 단계(30)는 생성된 학습 문장 집합을 사용하여 학습에 사용할 자질을 추출하는 자질 추출 과정(310)과; 추출된 자질을 사용하여 학습하고 입력된 문서에 범주를 할당하는 문서 범주화 과정(320)을 포함한다. 자질 추출 과정(310)에서는 카이 제곱 통계량(χ^2 statistics) 값을 사용하고, 문서 범주화 과 정(320)에서는 문서 분류기로서 베이시안 확률 모델(Bayesian Probability Model)을 사용한다.
도 2는 수집된 문서 집합의 문서 정규화 과정과 문장 단위 분할 과정을 거친 후에 언어 분석과 태깅 과정을 통해 각 문장의 내용이나 특징을 잘 반영하는 내용어를 추출하는 과정을 예시한다.
먼저 수집된 문서 집합은 문서 정규화 과정을 통해 한자어나 각종 태그 등을 제거하고 문장단위로 분할된다(S11). 분할된 문장은 예시된 바와 같이 형태소 분석 및 태깅을 통해 언어적, 통계적 분석을 통해 각 형태소 별로 품사를 결정한다(S12).
품사 중에 문장의 특징을 잘 나타내는 품사인 명사(외래어 포함)와 동사만의 내용어를 추출한다(S13). 여기서 추출된 내용어 중에는 여러 문장에서 공통적으로 많이 나타나기 때문에 문장의 내용을 구분하기 위해 별다른 정보를 주지 못하는 불용어들이 있다. S13의 예에서 '기본[명사]'이 불용어에 해당하는데 이를 제거하기 위해 미리 불용어에 대한 사전을 구축해서 사전에 등록되어 있는 단어는 제거하여 최종적으로 해당 문장의 내용어를 추출한다(S14).
도 3은 문장 집합으로부터 각 범주별 학습 문장 집합을 자동으로 생성해내는 과정을 예시한다. 수집된 문서 집합의 문장 집합이 S21과 같고 범주별 핵심어가 S22와 같으며 '음악'과 '인터넷'이라는 두가지 범주가 있다고 가정하자. '음악' 범주의 핵심어인 '음악'을 내용어로 가지고 있는 문장 1는 '음악'범주의 대표 문장으로 추출되고(S23), '인터넷' 범주의 핵심어인 '인터넷'을 내용어로 직접 가지고 있 는 문장 2은 '인터넷' 범주의 대표 문장으로 추출된다(S24). 범주별 핵심어를 직접 내용어로 가지지 못하는 문장은 미 분류 문장으로 분류된다(S25).
추출된 대표 문장들만으로 각 범주의 학습을 위한 학습 문장 집합이 되기에는 양이 부족하기 때문에, 대표 문장으로 추출되지 못한 미 분류 문장들과 각 범주의 대표 문장과의 유사도 측정을 통해 가장 유사도 값이 높게 나오는 범주로 미 분류 문장을 할당시킨다(S26). 문장 3과 문장 4는 핵심어를 가지고 있지 않기 때문에 미 분류 문장으로 분류되었으나 유사도 측정 과정(S26)을 거쳐 문장 3은 '음악' 범주에 할당되고(S27), 문장 4는 유사도 측정의 값이 어느 한계값 이상이 되지 않으므로 어느 범주에도 해당되지 않는 것으로 간주되어 계속 미 분류 문장 집합에 속하게 되고 결국 학습에 참여하지 않는다(S28).
본 발명에서는 문장간 유사도 측정 방법이 매우 중요한데 이를 위해 기존에 정보 검색에서 사용하는 단순한 방법들을 사용하지 않고 도 4와 같이 단어 유사도 행렬(S41)과 문장 유사도 행렬(S42)을 사용하여 반복 계산하고 문장간 유사도를 계산한다. 유사한 단어는 유사한 문맥에 위치하는 경향이 있으므로 이를 이용하여 문맥 정보를 반영하여 문장간 유사도를 측정한다. 본 발명에서 단어와 문장은 상호 보충적인 역할을 수행하는데, 문장은 포함하고 있는 단어들에 의해 표현되고, 단어는 그 단어를 포함하고 있는 문장들에 의해 표현된다. 즉, 문장은 유사한 단어들을 많이 포함할수록 유사한 문장이고 단어는 유사한 문장에서 많이 사용될수록 유사한 단어이다. 이를 반영하기 위해 2개의 유사도 행렬(S41),(S42)을 구성하고 반복 계산을 통해 계산된 유사도의 값이 서로에게 반영되도록 하였다.
단어 유사도 행렬(S41)의 행과 열은 유사도를 측정하고자 하는 범주별 대표 문장과 미 분류 문장들에 포함되어 있는 모든 내용어들로 구성되어 내용어 사이의 유사도 값을 가지며, 문장 유사도 행렬(S42)은 대표 문장과 미 분류 문장들의 유사도 값을 나타내게 된다.

발명의 효과

본 발명은 수작업에 의해 범주가 할당된 대량의 학습 문서 생성 작업 없이 문서 범주화를 수행하게 함으로써 적은 비용으로 문서 범주화를 수행하고자 하는 온 라인 상의 문서 범주화 응용 영역에서 유용하게 사용할 수 있는 효과가 있

으며, 또한, 대량의 학습 문서 생성 작업에 본 발명에서 제안된 기법을 사용한다면 작업에 소요되는 많은 시간과 인력을 최소화하여 학습 문서를 생성할 수 있는 효과가 있다.

(57) 청구의 범위

청구항 1.
인터넷에서 수집된 문서의 자동 문서 범주화 방법에서,
수집된 문서를 정규화하고, 상기 정규화된 문서를 문장 단위로 분할하며, 상기 분할된 문장 단위를 언어적 분석을 통해 각 문장의 내용어를 추출하는 단계; 및
상기 정규화, 분할화 및 추출화된 문장 단위의 집합에서 대표 문장을 추출하고, 단어 유사도 행렬과 문장 유사도 행렬을 사용하여 상기 대표 문장과 상기 문장 단위의 유사도를 측정하여 각 범주별로 분류하는 것에 의해 학습 문장을 생성하는 단계를 포함하는 것을 특징으로 하는 자동 문서 범주화 방법.

청구항 2.
(a)수집된 문서를 문장 단위로 분할하고 형태소 분석하여 내용어를 추출하는 단계;
(b)입력된 핵심어를 이용하여 각 범주의 대표 문장을 추출하는 단계;
(c)상기 추출된 대표 문장이 각 범주의 특성을 잘 나타내고 있는지를 검증하여 순위화하는 단계;
(d)상기 추출된 대표 문장과 대표 문장으로 추출되지 못한 미 분류 문장과의 문장간 유사도 측정을 통하여 학습에 사용될 학습 문장 집합을 생성하는 단계; 및
(e)상기 생성된 학습 문장 집합을 사용하여 학습에 사용할 자질을 추출하고 학습하여 문서에 범주를 할당하는 단계를 포함하는 비지도 방식의 자동 문서 범주화 방법.

청구항 3.
제 2항에 있어서, 상기 수집된 문서의 내용어 추출단계는 수집된 문서를 기계적 처리가 가능하게 하는 문서 정규화 단계와; 정규화된 문서의 문장을 문장 단 위로 분할하는 문장 분할 단계와; 분할된 문장의 형태소 분석 및 태깅 단계를 포함하고, 내용어의 추출은 불용어 사전을 사용하는 것을 특징으로 하는 자동 문서 범주화 방법.

청구항 4.
제 2항에 있어서, 상기 대표 문장을 추출하는 단계는 입력된 범주별 핵심어가 내용어로 직접 포함되어 있는 문장들을 추출하고 이들을 각 범주의 특성을 가장 잘 나타내는 문장으로 간주하는 것을 특징으로 하는 자동 문서 범주화 방법.

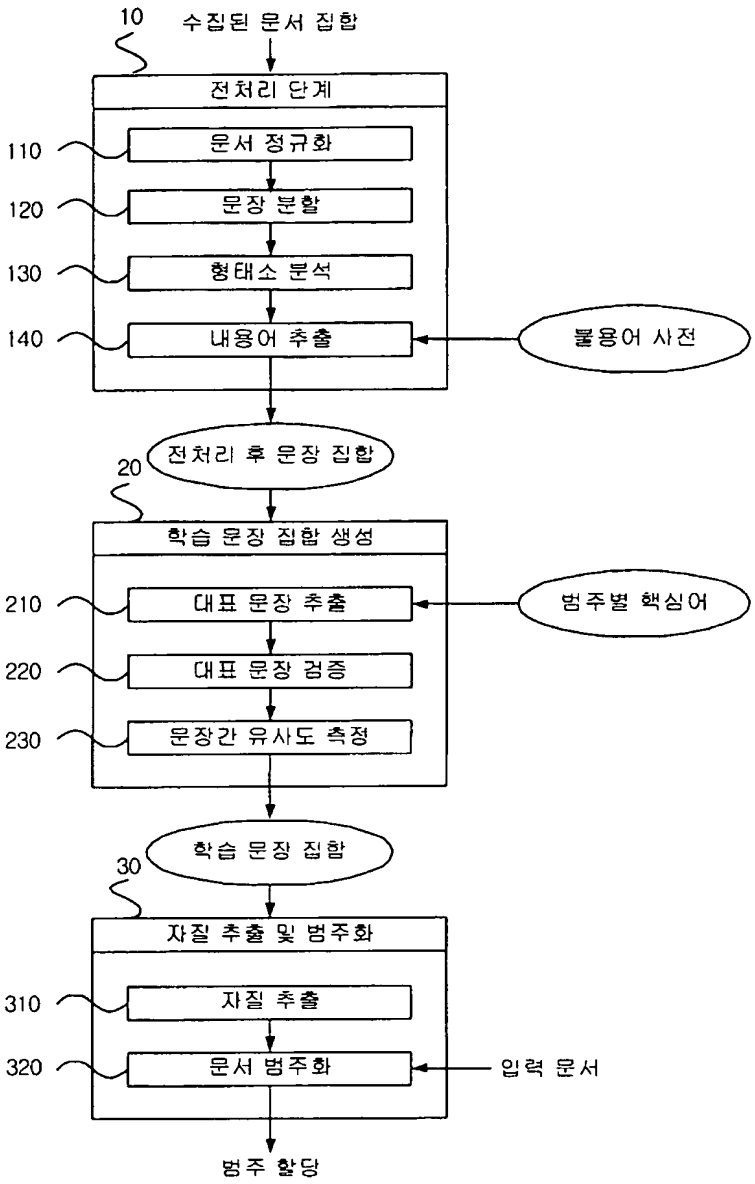
청구항 5.
제 2항에 있어서, 상기 대표 문장 검증 및 순위화 단계는 용어 빈도(TF)와 역범주 빈도(ICF)를 사용하여 추출된 대표 문장의 각 내용어에 가중치를 부여하는 단계를 포함하는 것을 특징으로 하는 자동 문서 범주화 방법.

청구항 6.
제 2항에 있어서, 상기 학습 문장 집합 생성단계에서 문장간 유사도 측정은 단어 유사도 행렬과 문장 유사도 행렬을 사용하여 반복 계산을 통해 얻어지는 것을특징으로 하는 자동 문서 범주화 방법.

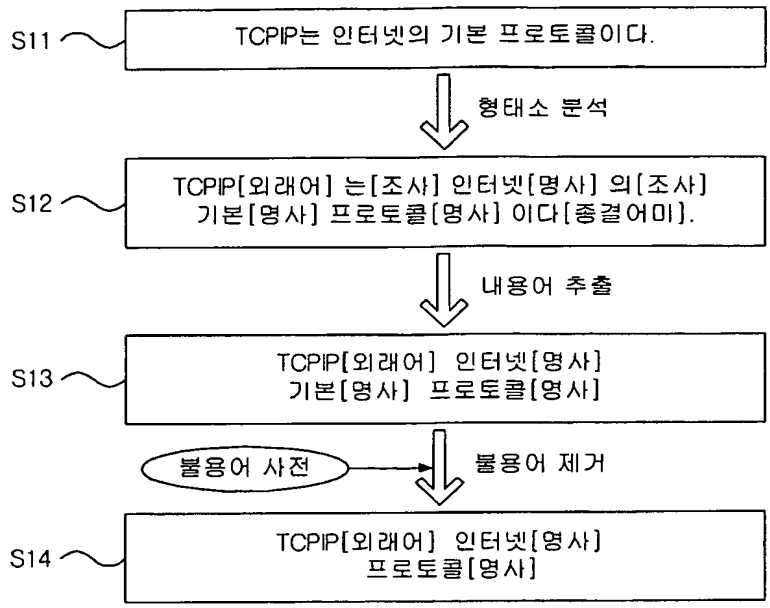
청구항 7.
제 6항에 있어서, 상기 단어 유사도 행렬의 행과 열은 유사도를 측정하고자 하는 범주별 대표 문장과 미 분류 문장들에 포함되어 있는 모든 내용어들로 구성되어 내용어 사이의 유사도 값을 가지며, 문장 유사도 행렬은 대표 문장과 미 분류 문장들의 유사도 값을 가지고 있는 것을 특징으로 하는 자동 문서 범주화 방법.

도면

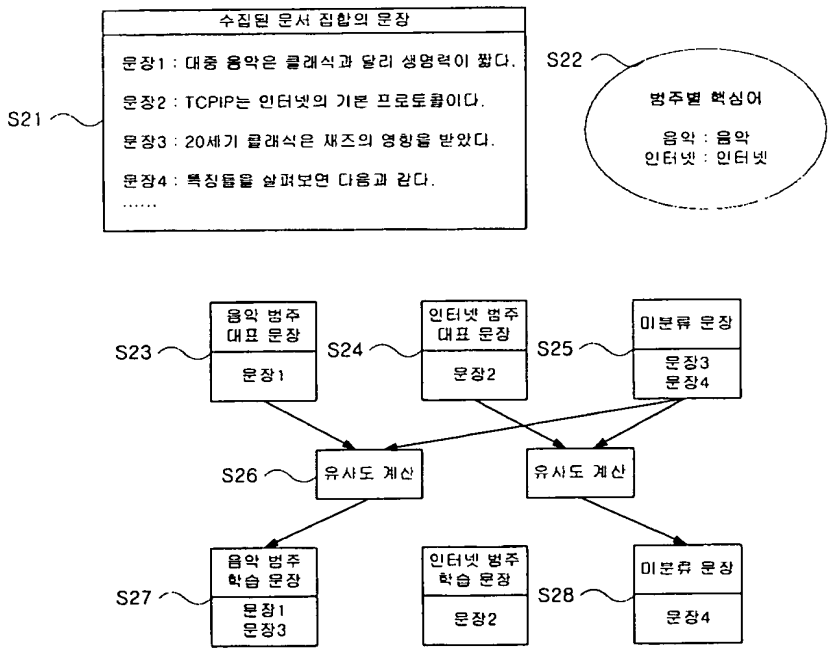
도면1


10
수집된 문서 집합
전처리 단계
110
문서 정규화
120
문장 분할
130
형태소 분석
140
내용어 추출
불용어 사전
전처리 후 문장 집합
20
학습 문장 집합 생성
210
대표 문장 추출
범주별 핵심어
220
대표 문장 검증
230
문장간 유사도 측정
학습 문장 집합
30
자질 추출 및 범주화
310
자질 추출
320
문서 범주화
입력 문서
범주 할당

도면2


S11
TCPIP는 인터넷의 기본 프로토콜이다.
형태소 분석
S12
TCPIP[외래어] 는[조사] 인터넷[명사] 의[조사]
기본[명사] 프로토콜[명사] 이다[종결어미].
내용어 추출
S13
TCPIP[외래어] 인터넷[명사]
기본[명사] 프로토콜[명사]
불용어 사전
불용어 제거
S14
TCPIP[외래어] 인터넷[명사]
프로토콜[명사]

도면3


수집된 문서 집합의 문장
문장1 : 대중 음악은 클래식과 달리 생명력이 짧다.
문장2 : TCPIP는 인터넷의 기본 프로토콜이다.
문장3 : 20세기 클래식은 재즈의 영향을 받았다.
문장4 : 특징들을 살펴보면 다음과 같다.
.....
S21
S22
범주별 핵심어
음악 : 음악
인터넷 : 인터넷
S23
음악 범주 대표 문장
문장1
S24
인터넷 범주 대표 문장
문장2
S25
미분류 문장
문장3
문장4
S26
유사도 계산
유사도 계산
S27
음악 범주 학습 문장
문장1
문장3
인터넷 범주 학습 문장
문장2
S28
미분류 문장
문장4

도면4